SEME 2017 : identification de véhicules en utilisant le numéro VIN

Rémi Besson, Christèle Etchegaray, Luca Ferrari, Samuel Nordmann

Rapport SEME : identification de véhicules en utilisant le numéro VIN

Rémi Besson, Christèle Etchegaray, Luca Ferrari, Samuel Nordmann

Table des matières

Résumé

Ce rapport présente le travail effectué dans le cadre de la SEME 2017 à Paris, et porte sur de l'analyse de données pour l'identification de véhicules dans le cadre de l'activité de l'entreprise Oscaro.

Remerciements

Les auteurs sont très reconnaissants envers les encadrants Oscaro du projet, Nils Grunwald, Stéphane Raux et Roland Thiolliere pour leur disponibilité et leur enthousiasme communicatif. Merci également aux organisateurs de la SEME, ainsi qu'à AMIES pour donner l'opportunité aux doctorants de découvrir d'autres environnements de recherche.

1 Introduction

1.1 L'entreprise Oscaro

L'entreprise Oscaro compte environ 600 employés, et son équipe de recherche se compose d'environ 10 personnes. Elle propose à des usagers (souvent des particuliers) la vente en ligne de pièces détachées de voiture.

Il existe entre 50 et 70 millions de types de véhicules sur le marché. Oscaro travaille avec différents fournisseurs de pièces, et collecte ainsi un stock de pièces détachées accompagnées d'informations sur les types de véhicules correspondants (une pièce pouvant servir à plusieurs types de véhicules). Il existe une classification hiérarchique des types de véhicules permettant de leur associer les bonnes pièces : du plus large au plus spécifique, il y a

1. le constructeur
2. le modèle
3. le type
4. le numéro Oscaro
5. le numéro Polk.

Lorsqu'un usager veut commander une pièce en ligne, il fournit des informations concernant son véhicule, afin de se voir proposer les pièces correspondantes. En France, il s'agit en général du numéro de plaque d'immatriculation. Oscaro doit alors lui associer un ou plusieurs éléments de la classification précédente afin de lui proposer les bons produits.

En France, la sous-préfecture dispose de la correspondance entre la plaque d'immatriculation et les informations contenues dans la carte grise, et cela suffit généralement pour identifier les pièces (moyennant une requête à la sous-préfecture facturée entre 5 et 20 centimes l'unité). Cependant, certains pays comme l'Allemagne ne fournissent pas ces informations, tandis que d'autres comme l'Espagne les fournissent mais disposent de bases de données peu fiables (données entrées à la main, de manière non systématique). Il s'agit donc ici d'utiliser un autre moyen de caractériser le véhicule.

1.2 Le Vehicle Identification Number

Depuis l'après-guerre, un numéro d'identification est gravé sur le chassis de chaque voiture : il s'agit du VIN (Vehicle Identification Number). Il est standardisé depuis 1983, et se compose de 17 caractères alphanumériques. Il peut être décomposé en trois parties (voir ci-dessous la figure 1) :

FIGURE 1 – Un exemple de VIN sur le chassis d'une voiture, séparé en trois parties : le WMI, le VDS, et le VIS.

Le code WMI (pour World Manufacturer Identifier) correspond aux trois premiers caractères, codant respectivement la zone géographique, le pays au sein de cette zone, et le constructeur. Ces informations son codées de façon explicite et normalisée dans le WMI.

Le code descripteur VDS correspond aux 6 caractères suivants. Il porte des informations sur le véhicule, sans être normé.

Le code indicateur VIS correspond aux 8 derniers caractères. Il n'est pas normé et est laissé à la discrétion du constructeur, avec comme seule contrainte que chaque véhicule doit avoir un code VIN unique. Ainsi, aucune règle ne peut être supposée quant à la manière dont un VIS peut coder les caractéristiques du véhicule : elle peut être différente selon les constructeurs, usines, années de production etc.

1.3 Problématique

Le problème qui nous a été proposé durant la SEME est de comprendre les liens entre le code VIN d'un véhicule et ses caractéristiques, c'est à dire d'établir des corrélations entre les codes VIN et la classification CMTOP (Constructeur-Modèle-Type-Oscaro-Polk).

Identifiant véhicule		**Classification pièces**
		Constructeur
	?	Modèle
VIN = WMI | VDS | VIS	$\longrightarrow$	Type
		Oscaro
		Polk

Plus précisement, notre objectif est de proposer un algorithme qui, à partir d'un code VIN, renvoit une liste réduite de caractéristiques possibles pour le véhicule. Cette liste devra être triée par pertinence. Une contrainte que doit vérifier cet algorithme est de réduire au maximum les "faux négatifs", c'est-à-dire que la liste proposée par l'algorithme doit toujours contenir les caractéristiques réelles du véhicule, quitte à devoir renvoyer une liste de plus grande taille (on pourra alors solliciter l'intervention de l'utilisateur pour discriminer le bon élément parmi ceux de la liste). La caractéristique "Polk" étant la plus précise, nous nous concentrerons surtout sur la liste des "Polks" possibles pour un véhicule à partir de son code VIN.

Dans un second temps, un autre objectif visait à identifier comment certaines caractéristiques secondaires du véhicules (boîte de vitesse, essence, etc) sont codées au sein du VIS. Cependant, nous n'avons pas abordé ce problème durant la semaine de travail.

Nous disposions d'une base de données renseignant les caractéristiques CMTOP ainsi que le code VIN de 30 millions de véhicules. Toutes ces données étaient anonymisées, et provenaient principalement de véhicules français. Du point de vue technique, nous avons utilisé le langage Python pour gérer les fichiers de données au format CSV.

Particularités de ce travail : il s'agit là d'un problème classique de classification supervisée mais sur une base de données très particulière. D'une part, le fait que les constructeurs utilisent des règles pour générer les codes VIN induit une importante structure dans la base de donnée, ce qui peut réduire la complexité des calculs de manière importante. D'autre part, ces règles étant inconnues et variables (en fonctions des constructeurs, unités de production etc.) on ne peut pas identifier cette structure a priori. Par conséquent, il est difficile de développer une intuition globale sur la base de données.

2 Identifier le constructeur : un dictionnaire

Une première analyse des données montre qu'à un WMI donné on associe systématiquement un identifiant de constructeur. Cette correspondance constitue quasiment une règle de constitution d'un VIN, au moins en France. Des contre-exemples existent néanmoins pour des véhicules étrangers, qui n'étaient pas représentés dans la base de données.

Il est donc possible ici de mettre en place en amont un algorithme de type dictionnaire pour utiliser directement cette correspondance sans avoir à faire appel à la base de données.

3 Une corrélation plus précise

En fixant WMI+VDS dans la base de données, on obtient un ensemble restreint de Polks possibles. Nous formulons alors l'hypothèse selon laquelle à un numéro VIN (extérieur à la base) donné, son Polk associé se trouve dans l'ensemble de Polks obtenu en filtrant la base de données par WMI+VDS.

Notons $\overline{\text{POLK}}$ l'ensemble des numéros Polk, $\overline{\text{VIN}}$ l'ensemble des numéros VIN, et notons $v = (v_1, v_2, v_3) \in \overline{\text{VIN}}$ un élément de cet espace, où chaque v_i correspond à une partie du VIN. Notons enfin $F : \overline{\text{VIN}} \to \overline{\text{POLK}}$ l'application qui à un numéro VIN associe son Polk : on a $P = F(v)$. Notre hypothèse se formule ainsi : soit $v^{\text{ext}} \in \overline{\text{VIN}}$ un numéro VIN inconnu. Alors,

$$F(v^{\text{ext}}) \in \{F(v) \mid v \in \overline{\text{VIN}} \text{ et } (v_1, v_2) = (v_1^{\text{ext}}, v_2^{\text{ext}})\}. \tag{1}$$

3.1 Validation de l'hypothèse par un test statistique

Nous décidons de tester cette hypothèse sur les données par validation croisée : nous partitionnons la base de données de manière aléatoire en un ensemble d'apprentissage (80% de la base) et un ensemble de test (20%). Nous validons l'hypothèse 1 avec un taux de succès de 99.75%. Nous considérons donc cette hypothèse comme valide dans ce qui suit.

3.2 Premières conclusions

L'hypothèse 1 ayant été validée, l'ensemble des Polks possibles pour un VIN donné ne contiendra que des Polks correspondant à des véhicules partageant le même WMI+VDS. Cela permet, dans un premier temps, de réduire de manière considérable la liste de Polk possibles pour un VIN donné. Le tableau 1 et la figure 2 donnent des informations chiffrées sur la taille de la liste réduite après ce premier tri.

Catégorie	**Constructeur**	**Modèle**	**Type**	**Oscaro**	**Polk**
Moyenne	1.02	1.18	1.59	1.69	1.95
Écart-type	0.18	0.81	3.81	4.09	6.32
Minimum	1	1	1	1	1
Q 25%	1	1	1	1	1
Q 50%	1	1	1	1	1
Q 75%	1	1	1	2	2
Maximum	26	79	180	198	506

TABLE 1 – Nombres moyens de valeurs différentes pour chaque catégorie à VIN1 et VIN2 fixés (89771 VIN3).

Dans plus de la moitié des cas, il ne reste qu'un seul Polk possible à l'issue de ce premier tri, et le véhicule de l'utilisateur est complètement identifié. Dans les trois quarts des cas, ce premier tri identifie complètement le "type" du véhicule, et nous laisse avec au plus deux Polks possibles. Cette méthode est donc satisfaisante du point de vue pratique.

Cependant, il existe quelques cas "pathologiques", pour lesquels le nombre de Polks possibles peut dépasser la vingtaine, et même être supérieur à 200. Dans ces cas là, cette première étape n'est pas suffisante et nous devons donc trouver un moyen de trier cette liste de "Polks". Pour cela, nous allons utiliser le code VIS, c'est à dire les 8 derniers caractères du code VIN, qui n'ont pas été pris en compte jusque là.

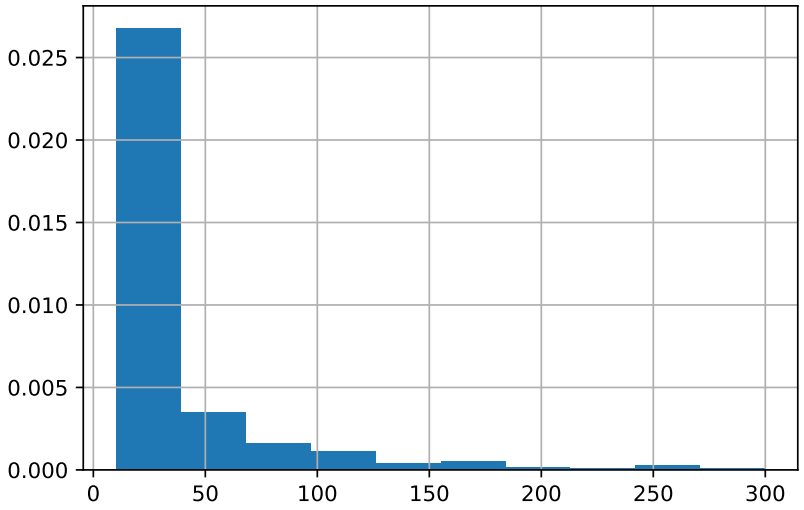

FIGURE 2 – Histogramme du nombre de Polk différents à WMI et VDS fixés.

4 Discriminer les identifiants grâce à une distance

Rappellons que notre objectif est d'utiliser la base de données pour être capable d'attribuer à un VIN extérieur un ensemble de numéros Polks possibles. Nous avons montré qu'il était possible d'utiliser les deux premières parties du VIN (WMI+VDS) pour se ramener à une base de données significativement réduite. Nous travaillons maintenant avec cette base réduite. Dans la suite, nous noterons $\overline{\text{VIN}}^r$ (respectivement $\overline{\text{POLK}}^r$) l'ensemble des numéros VINs (respectivement POLKs) représentés.

Il s'agit d'associer à chaque Polk y figurant un *score de pertinence* permettant de discriminer les Polks à proposer en priorité à l'utilisateur.

Pour cela, nous proposons d'effectuer sur la base réduite la démarche suivante :

— définir une distance entre les parties terminales du VIN (les VIS) de façon à ce que des VIS proches pour cette distance aient le même Polk.

— attribuer à chaque Polk de la base de données réduite un score de pertinence défini à partir de la distance entre les VIS de la base associés à ce Polk et le VIS extérieur.

Il suffit ensuite de proposer à l'utilisateur soit directement la liste hiérarchisée des Polks possibles, soit une liste réduite grâce à un critère à définir.

4.1 Définition d'une distance entre VIS.

Le VIS est une succession de 8 caractères alphanumériques. Pour définir une "distance" entre deux VIS, une première idée serait de compter le nombre de caractères différents : il s'agit de la **distance de Hamming**. Soient u_3, v_3 deux VIS :

$$d(u_3, v_3) = \sum_{k=1}^{8} \mathbb{1}_{\{u_3(k) \neq v_3(k)\}}.$$

Cependant, il semblerait que certaines parties du VIS soient plus importantes que d'autres pour déterminer le Polk, certaines parties étant probablement totalement arbitraires. D'où l'idée

de pondérer chaque caractère en fonction de son importance :

$$d^w(u_3, v_3) = \sum_{k=1}^{8} w_k \mathbb{1}_{\{u_3(k) \neq v_3(k)\}}.$$

La fonction distance que nous cherchons à définir dépend ainsi de 8 paramètres $w = (w_i)_{i \in \{1,\dots,8\}}$, qu'il s'agit de choisir de manière optimale à partir de notre base de données réduite.

Remark 4.1. *Vers une discrimination plus précise. Nous choisissons donc ici de définir une distance pour chaque classe d'éléments partageant le même WMI+VDS, dans le cas où la base de donnée réduite obtenue est encore assez riche.*

Rappelons que la manière de coder le VIS peut différer en fonction des constructeurs, unités de productions, années etc. Nous pourrions donc également envisager de choisir les poids en utilisant une sous-classe de la base différente. Il faudrait alors réaliser un compromis entre la précision des informations gagnée en restreignant la base, et la baisse de pertinence statistique conséquente.

Choix du vecteur de poids w.

Soit X_k le k-ème caractère du VIS, et Polk $\in \overline{\text{POLK}}^r$ un numéro Polk donné. La quantité $\mathbb{P}[X_k = \ell \mid \text{Polk}]$ est la probabilité d'apparation du caractère ℓ en k-ème position du VIS, pour ce Polk. Nous estimons cette probabilité par la fréquence correspondante dans notre base de données. Nous définissons l'*entropie* du k-ème caractère :

$$H(X_k \mid \text{Polk}) = -\sum_{\ell} \mathbb{P}[X_k = \ell \mid \text{Polk}] \log(\mathbb{P}[X_k = \ell \mid \text{Polk}]).$$

L'entropie est une fonction très utilisée en théorie de l'information. Formellement, l'entropie mesure la "quantité d'information" contenue dans une variable aléatoire.

Ici, $H(X_k \mid \text{Polk})$ mesure la quantité d'information sur le Polk contenue dans le k-ème caractère du VIS. En effet, moins il y aura de variabilité pour le k-ème caractère du VIS, plus le k-ème caractère contient de l'information sur le Polk, et plus l'entropie sera proche de 0. Par exemple, pour un Polk fixé, si le 2-ème caractère du VIS est toujours un "A", alors la variabilité du 2-ème caractère est nulle, et l'entropie sera nulle également.

Cependant, cette définition ne tient pas compte de la quantité d'information *totale* contenue dans le k-ème caractère :

$$H(X_k) = -\sum_{\ell} \mathbb{P}[X_k = \ell] \log(\mathbb{P}[X_k = \ell]).$$

En effet, si la quantité $H(X_k)$ est proche de 0, cela signifie qu'indépendemment du Polk, le k-ème caractère a une faible variabilité. La quantité $H(X_k \mid \text{Polk})$ est alors très petite également, mais cela ne signifie pas que le k-ème caractère contient beaucoup d'information sur le Polk : il faut comparer $H(X_k \mid \text{Polk})$ avec $H(X_k)$.

Nous définissons alors *l'information mutuelle* :

$$I(X_k, \text{Polk}) = H(X_k) - H(X_k \mid \text{Polk}).$$

L'information mutuelle $I(X_k, \text{Polk})$ mesure la quantité d'information sur le Polk donné apportée en moyenne par la connaissance de X_k. Formellement, elle correspond à la quantité d'information réelle sur le Polk contenue dans le k-ème caractère du VIS. Ainsi plus cette quantité est grande plus le k-ème caractère est caractéristique du Polk considéré.

Nous devons à présent définir les poids w_k en se basant sur l'heuristique selon laquelle plus l'information mutuelle est grande, plus le k-ème caractère devra avoir de l'importance. Ainsi, nous posons

$$w_k = a \sum_{\text{Polk} \in \overline{\text{POLK}}^r} I(X_k, \text{Polk})$$

où $a > 0$ est un paramètre de sensibilité.

4.2 Score de proximité pour le Polk

Rappelons que nous avons défini une distance permettant de mesurer la proximité de deux VIS donnés pour l'encodage du numéro Polk. Il s'agit maintenant d'utiliser cette distance pour associer à chaque Polk de la base de données réduite un score de proximité avec le véhicule à identifier. Nous savons que les VINS présents ont en commun leurs deux premières composantes (WMI et VDS). Nous notons alors $\overline{\text{VIN}}_3^r$ l'ensemble des VIS représentés, et définissons l'application surjective

$$\begin{aligned} G : \overline{\text{VIN}}_3^r &\to \overline{\text{POLK}}^r \\ v_3 &\mapsto G(v_3) \end{aligned}$$

qui à un VIS donné associe son Polk.

A un numéro VIN extérieur v^{ext} donné, nous associons une base de données réduite et une distance d^w. Pour tout Polk présent dans la base réduite, nous associons un score de proximité à v^{ext}, qui correspond à la moyenne arithmétique des distances entre tout VIS associé à ce Polk et le VIS extérieur v_3^{ext}. Plus précisément, nous écrivons

$$\text{Score}(\text{Polk}) = \frac{1}{|G^{-1}(\text{Polk})|} \sum_{u_3 \in G^{-1}(\text{Polk})} d^w(v_3^{\text{ext}}, u_3)\,.$$

Le Polks ayant les scores les plus bas sont ainsi les plus susceptibles de correspondre au VIN extérieur donné.

Cette méthode peut être sujette à de nombreuses critiques. Notamment elle ne permet pas d'exclure les Polks représentés par trop peu d'élements. On pourrait envisager des alternatives, comme l'algorithmes des k plus proche voisins. Des tests quantitatifs pourraient alors être menés afin de comparer les différentes méthodes.

4.3 Tests numériques

4.3.1 Etudier la distribution des caractères du VIN à la lumière de l'entropie.

Il peut être intéressant tout d'abord d'étudier, à constructeur fixé, l'entropie moyenne observée pour chaque caractère du VIN. Les résultats sont présentés sur la figure 3. Nous avons, à constructeur fixé, calculé l'entropie de chaque caractère du VIN. Comme attendu nous trouvons bien une entropie nulle, ou très proche de 0, pour les premiers caractères du VIN. A l'inverse les derniers caractères du VIN présentent une entropie élevée, supérieure à 2. Il est à noter que la valeur maximale pouvant être prise par l'entropie d'une variable aléatoire définit dans $\{0, 1, ..., 9\}$ est atteinte par la distribution uniforme sur $\{0, 1, ..., 9\}$. L'entropie vaut alors : $\sum_{k=0}^{9} -1/10 \times \log(1/10) = log(10) \approx 2.3$. Les derniers caractères du VIN ont donc une entropie très élevée, proche de celle qu'aurait une variable remplie de manière totalement aléatoire (uniforme).

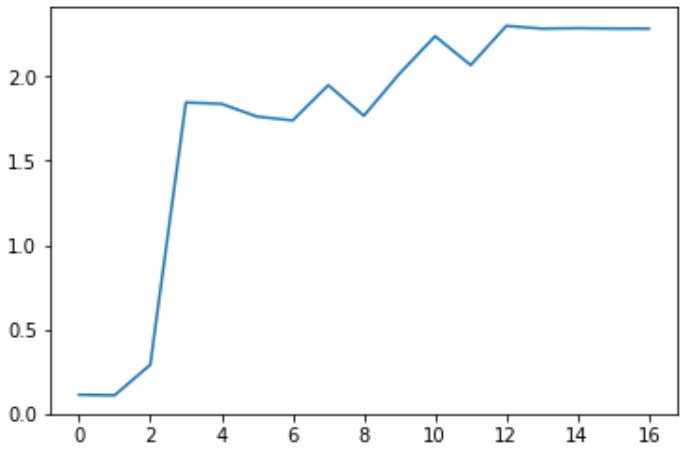

FIGURE 3 – Entropie des 16 caractères du VIN pour un **Constructeur** fixé.

4.3.2 Les résultats de l'algorithme de recommandation de VIN.

Nous avons mis en oeuvre l'algorithme d'apprentissage présenté sur un ensemble de données contenant approximativement 26 700 000 éléments. Cet ensemble a été partitionné en un ensemble d'apprentissage et un ensemble de test. L'ensemble d'apprentissage est destiné à entraîner les poids des différentes distances. Puis, nous testons l'algorithme résultant sur l'ensemble de test : chaque élément nous fournit un nouveau VIN, auquel l'algorithme associe une liste de Polks possibles ordonnés par score de pertinence. Nous pouvons ainsi représenter les résultats sous deux formes.

La figure 4 représente l'histogramme du rang obtenu pour le Polk correct pour chaque test. La figure 5 représente le rang obtenu pour le Polk correct en fonction du nombre de Polks possibles dans $\overline{\text{VIN}}^r$.

Les résultats ont été ici obtenus pour une base de test dont la taille est de 0.01% la taille de la base de données, soit environ 2670 éléments. Il serait évidemment intéressant d'effectuer des tests plus nombreux, mais ces résultats constituent une première indication satisfaisante de l'efficacité de notre algorithme.

5 Une alternative : les réseaux de neurones.

Pour répondre à notre problème de classification supervisée, une approche concurrente à celle que nous avons présentée plus haut, consiste à apprendre un réseau de neurone prenant en entrée un VIN et donnant la probabilité d'appartenance à chaque Polk.

On travaille ici à WMI fixé, il y a donc un nombre connu de Polk possibles. Il s'agit d'entraîner un réseau pour chaque WMI.

Cette approche comporte plusieurs difficultés. Nous nous retrouverons ainsi parfois à apprendre un réseau de neurone à partir d'un faible nombre de données (lorsque le WMI en question est peu fréquent dans la base). De plus optimiser un réseau de neuronne n'est pas toujours chose aisée et de nombreux hyperparamètres doivent être fixés, apprendre un réseau pour chaque sous-tâche est, de ce point de vue, une difficulté supplémentaire.

Cette piste n'a pas été explorée au cours de la SEME mais constitue une direction pertinente pour des travaux futurs.

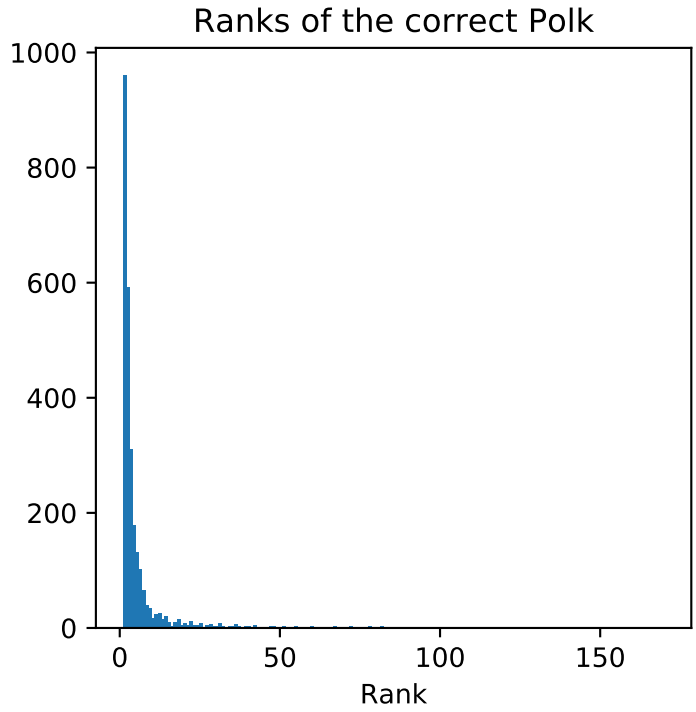

FIGURE 4 – Histogram of the rank of the correct Polk.

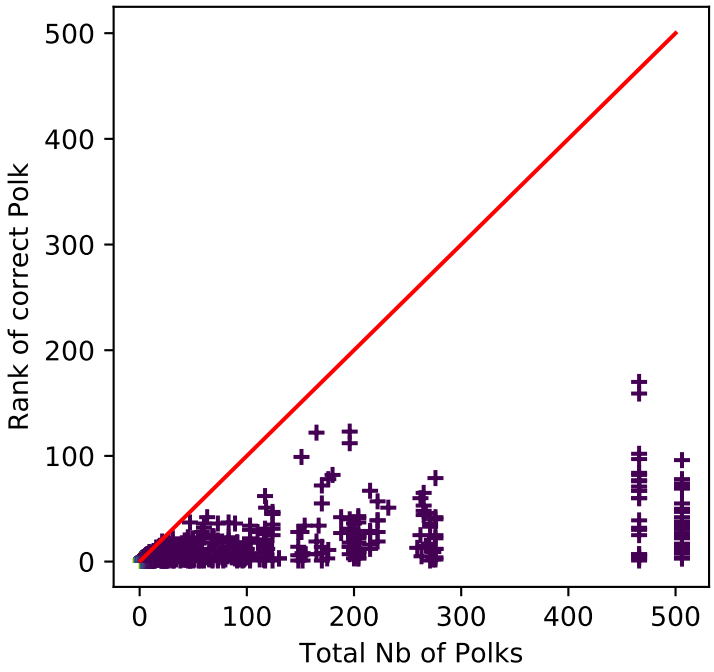

FIGURE 5 – Scatter plot of the rank of the correct Polk as a function of the number of possible Polks.